Scientific Data Management: Progress Since 2010 Conference IWGDD/CENDI/EPA

Lynne Petterson, PhD
Office of Science Information Management, ORD
United States Environmental Protection Agency

CENDI
March 8, 2012

United States
Environmental Protection
Agency

Why now?

- Data deluge: Resources unable to keep up
  - Lots of existing data & growing exponentially
    - Network storage: 90 TB
    - HPC ASM: 4.1 PB (up from 580 TB six years ago)
  - Exponential growth
    - Network storage growth: 60-70% / year
    - HPC appx 145 TB / month
- Data inefficiencies
  - Duplication & unnecessary retention
  - Data inaccessible, unusable, or simply can’t be found
- Barriers to collaboration
  - Difficult to identify potential collaborators
  - Unwillingness to share data
  - Inconsistent SDM practices across ORD

“There’s a lot out there, but I can’t find any of it”.

“Data tsunami any day now”

“It’s easier to redo the science than to access the old data.”

“The data issue is growing faster than the solution.”

Why Else?

- Federal government focus on digital data
  - The America COMPETES Reauthorization Act of 2010, Section 103 (under the Office of Science and Technology Policy) entrusts Federal agencies to develop “policies for the management and use of Federal scientific collections to improve the quality, organization, access – including online access, and long-term preservation of such collections for the benefit of the scientific enterprise.”

- Recent efforts and initiatives
  - IWGDD/CENDI/EPA multi-agency workshop on SDM (2010)
  - *ORD’s* IM/IT *Strategic Action Plan* (2010)
  - ORD’s collaborative Path Forward efforts
  - Other Agency efforts
  - Draft EPA Data Policy supporting data access, discovery, sharing under data.gov

“I think it’s the perfect time… There’s a lot big changes coming with the Path Forward. If we change how collect or manage data at the same time, it’s congruent.”

EPA
United States
Environmental Protection
Agency

First Steps

- Identified how scientific data are managed elsewhere (EPA & outside); Co-sponsored IWGDD/CENDI/EPA Conference (5-2010)

- Convened an ORD workgroup to develop a DRAFT SDM Policy
  - ORD policy is to manage scientific data in a manner that: 1) recognizes scientific data as an asset; 2) considers data value and costs; and 3) is consistent across the organization.

- Created draft high-level procedural guidance
  - Develop a Scientific Data Management Plan that covers the full data life cycle
  - Identify scientific data with metadata
  - Identify data storage needs
  - Manage scientific data for appropriate sharing and access control
  - Manage scientific data for organization and control
  - Ensure scientific data knowledge is captured and retained
  - Retain data according to records management requirements
  - Enable reuse of scientific data

Workgroup Members

- **NRMRL:** Joel Allen/Cincinnati, Dave Burden/Ada, Sue Kimbrough/RTP, Leisha Vance/Cincinnati
- **NCCT:** Steve Little/RTP
- **NHSRC:** Eletha Roberts/Cincinnati
- **NCEA:** Deborah Wales/RTP, Lyle Burgoon/RTP
- **NCER:** Alva Daniels/DC, Lisa Doucet/DC
- **OSP:** Mimi Dannell
- **NHEERL:** Steve Edwards/RTP, Chris Russom/Duluth, Doug Lothenbach/Duluth, Dave Bolgrien/Duluth, Tony Olsen/Corvallis, Jim Harvey/Gulf Breeze, Tom Hollenhorst/Duluth, Russell Kreis/Grosse Ile, Jeff Hollister/Narragansett
- **NERL:** Ann Pitchford/LV, John Martinson/Cincinnati, Kurt Wolfe/Athens, Myriam Medina-Vera/RTP, Eunice Varughese/Cincinnati, Carry Croghan/RTP, Michelle Henderson/Cincinnati, Ken Schere/RTP
- **ORMA:** Peter Evanko/DC, Deborah Heckman/DC
- **OSA:** Lara Autry/RTP, Michael Bender/DC
- *OSIM and OEI Ex Officio Participation*

EPA
United States
Environmental Protection
Agency

Refine Guidance

- Reviewed OSIM ‘Data Study’
  - What data exist, where, how much, growth rates, etc
- Obtained input for SDM guidance development
  - June-Early October 2011
  - ~60 face-to-face and telephone interviews with ORD staff
    - Represent L/C/Os, disciplines, and roles
    - 12 central and remote locations
    - Gathered best practices, illustrative examples, and potential challenges
  - Findings are being analyzed and incorporated into procedural guidance
  - Draft procedural guidance anticipated in Q2 (Feb-March) 2012
  - Procedural guidance will ultimately serve as a desk reference

> “Groups need to come together to share this load. Otherwise it won’t happen.”

United States
Environmental Protection
Agency

Who We Interviewed

(Total number of interviews = 64)

L/C/O		Location		Discipline		Role	
NCCT	2	Ada	2	Risk Assess	5	Researcher	41
NCEA	7	Athens	1	Statistics	3	QA	10
NERL	11	Cincinnati	18	Genomics	3	Science Manager	7
NHEERL	25	Corvallis	1	Modeling	14	Data Manager	4
NHSRC	3	DC	3	GIS	5	Records	4
NRMRL	14	Duluth	3	Lab Science	15	Scientific Computing	1
OSA	1	Grosse Ile	1	Monitoring	11		
OEI	1	Gulf Breeze	13				
		Las Vegas	1				
		RTP	21				

Themes from Interviews

- ORD scientists are managing data, but there are inconsistencies and variability
- QAPP already requires some data management, but does not provide specific guidance with regard to SDM
- Confusion exists about current SDM requirements, including:
  - Records Management
  - Employee departures
  - Multi-L/C/O teams
- Scientists are asking for guidance & training
- Corroborates findings from *ORD's IM/IT Strategic Action Plan* (2011)

"The data issue is growing faster than the solution."

Principles for SDM Implementation

- Do no harm
  - Start small
  - Add detail and structured guidance to current requirements
- Lean on me
  - OSIM initially does the heavy lifting
  - OSIM establishes framework for minimizing effort for the scientists
- Achieve initial results quickly
  - Start with straightforward issues

"It's easier to redo the science than to access the old data."

United States
Environmental Protection
Agency

Next Steps: Early Adopters

- OSIM collaboratively creates SDM plans for “Early Adopters”
  - Steve Edwards: Biomarkers (NHEERL)
  - Sue Kimbrough: Near Road (NRMRL)
  - Anne Neale: National Atlas (NERL)
  - Linda Harwell: Sustainability Indicators (NHEERL)

- Anticipated results include:
  - Demonstrate benefits of SDM planning, such as:
    - Saves time and effort in the long run
    - Supports collaborative teams
    - Supports management requirements for ORD research portfolio
  - Build consensus from the ground up
    - Provides input to SDM process
    - Demonstrates feasibility and relative ease of SDM
    - Encourages culture change

> “Scientists are data generators and users, but NOT managers of large sets of data. They can’t do everything. But if you can’t trust your data, and your data management practices aren’t sound, you’ll fail an audit, make faulty assumptions in your data, fail the agency. “

United States Environmental Protection Agency

Concurrent SDM Tasks

- Records management
  - Add consistency to existing science data records schedules
  - Develop and implement guidance and training in conjunction with ORD Records Lead

"This isn't dictating, it's providing a structure that people should have been following all along."

- Employee departures
  - Develop and implement a consistent process for when people leave or retire
    - Document names and locations of files
    - Move data from C drive to network
- Contracts language for extramural data
  - First: Contracts
  - Later: Cooperative Agreements, Grants, IAGs

"I've been screaming for a long time that we need some sort of guidance. I'm willing to do anything, just tell me what I need to do."

- OSIM SDM support
  - Directly from IMSD to researchers
  - Provide SDM contract vehicles to ORD researchers

Input from Gulf Breeze

- Tools and Scientific Applications
  - "No one knows the tools/applications that are out there" (in ORD)
  - What tools are available, what data are created, and how to access both

- "Found" data (ie 'secondary data')
  - "How decide which interim geospatial coverages to keep?" (Early Adopter..)
  - Need for a tool to use and manage references (endnote + endnote +…)
  - Desperately seeking an ORD repository for large, final data sets

- OSIM Role (help 'comply w/ guidance' versus 'manage the data')
  - Comply with guidance
    - Direct support from OSIM (eg creating SDM Plans) ("boots on the ground")
  - Day-to-day management of scientific data
    - Request for OSIM-funded FTEs to manage data
    - Potential OSIM BPA for managing scientific data

"My first reaction is that it takes only about an hour to type all the stuff identified in the SDM Plan. Compiling the information may take longer…"

EPA United States Environmental Protection Agency

More Input from GB

“I’m not doing it until someone makes me do it.”

- SDM Plan Review and Approval
  - Who reviews the SDM Plans? QAM? Branch Chiefs? What about multi L/C/O?
- SDM Plans as “Living Documents”
  - Recognition that “SDM plans will be neither perfect nor complete initially” (evolving)
  - Complexity of SDM Plans likely to depend on Research Category (1,2,3,4…)
- Storage Charges
  - Perception of enterprise network storage as prohibitively expensive
    - External hard drives without backup
- Worthy future explorations
  - “All data from contractor must be in paper form so it can be sent to NARA”
  - “Do single, standalone georeferenced data points require full FGDC metadata? “

United States Environmental Protection Agency

Still More Input & Quotes

- ORD needs to ensure accessibility and awareness
  - Need communication and outreach re/ data
  - Need a place to showcase the data

"I have a map on the NOAA website since there is no place to put it on an EPA site."

"I'm not doing it until someone makes me do it."

" Who will baby sit models, systems, and data over time? Science can get stale. Think about the data life cycle at the beginning and eventual sunsetting."

"Information management begins with planning. What you do now impacts what you do later."

"People should be thinking about SDM issues in their research plans. They've ignored it until now."

"DM is an integral piece, but not the only piece, of managing your science." "

"Data Managers must be part of the research teams."

"Unless there's buy in from Branch Chiefs, there won't be any data management."

"The more outside attribution you have for your research, the greater your promotion potential."

"QA and SDM play roles in making data and research scientifically defensible."